Information Visualization for Human-Centered Data Analysis

Information visualization gives us the opportunity of synergizing the promise of the big data deluge with the high bandwidth of the human perception system, by creating actionable insight for analysts. This is inherently challenging, however, as despite the ever-increasing data size and complexity, the available screen real estate for visualizations remains mostly limited. Therefore, one of the key problems is to control the resulting information loss for preserving fidelity of the data and for faithfully expressing the features in the visualization. The information visualization literature has traditionally lacked models for quantifying different visual features and structures, which would help adapt the visual representations for maximizing user benefit. Another non-trivial problem is that these visual representations have to be intuitive enough for simplifying the sense-making process of users, especially for visualization non-experts like domain scientists, journalists, business analysts, etc. However, compared to our ability to generate more data, our understanding of human perception system is still limited, leading to a lack of standardized criteria for generating effective visualizations.

To address these problems, my research interests have centered on bridging the system-side of visualization with the human side, by devising both *quantitative* and *qualitative* methods. On the quantitative side, I have developed a closed loop model for measuring visual information content with the help of perceptually motivated metrics, and for subsequently optimizing visual encodings for supporting users' analytical tasks. I have studied how the metrics help in controlling user interaction and visual search for patterns, in real-world scenarios like high dimensional, temporal data visualization and privacy-preserving visualization. On the qualitative side, I have collaborated with climate scientists for understanding and evaluating how visualization design, in both static and interactive form, can be better adapted to suit their data analysis needs.
By gaining this dual perspective on visualizations, from the designer's side and the domain experts' side, I have been able to better reflect on how to augment visualization tools with features for exploiting our visual thinking capabilities, and thereby solve problems related to large, complex, and heterogeneous data.

Research Focus

In course of my doctoral dissertation research and then as a postdoc, I have worked on various research projects spanning a number of areas: for instance, building perception-based visualization models and taxonomy, applying those models for high-dimensional data analysis and privacy-preserving visualization; building visualization systems for biologists, climate scientists; and evaluating and improving existing visualization approaches by climate scientists. These different projects can broadly be grouped under three distinct but related threads of research: i) privacy-preserving visualization, ii) visualization approaches for climate data analysis, and iii) visual uncertainty models. Below, I highlight the key aspects of these research initiatives, including their motivation and the planned work.

Privacy-preserving visualization

One of the key contributions of my dissertation thesis was a technique and framework for privacy-preserving visualization. In fact, my advisor, Dr. Robert Kosara, and I were the first ones to propose a privacy-preserving visualization technique [5]. Till then, the issue of how to deal with privacy in handling sensitive data had been ignored the visualization community. In our first project, we used existing anonymization techniques used by the data mining community and used their output for visualizing sensitive data. We found that approach led to unusable visual representations due to the discrepancy between the data space and the screen-space.

Further work led to adapting existing work on data anonymization for developing visualization-based privacy metrics. The notion of privacy-preserving visualization opened a rather counter-intuitive research question: how to intentionally

*Numerical citations refer to the publications mentioned in my CV.

hide information in visualization and how much information loss is sufficient? A corollary to this is to study the amount of degradation of utility that is a consequence of privacy-preservation.

To address these questions, in collaboration with Dr. Min Chen (Oxford University) we proposed a model for quantifying the level of privacy and the loss of utility in cluster-based visualizations [1]. We further extended this work by taking attack scenarios into account and proposing how visualizations can be optimized for protection against manipulation by attackers. Papers related to handling attack scenarios and optimization of cluster-based visualizations are currently under review [9, 10].

Going forward, I would like to continue running experiments for testing the robustness of the privacy-preserving model in response to attack-scenarios. The idea of privacy-preserving visualization is novel, and I want to build infrastructures that can support this idea, for instance, a client-server model where we identify two distinct group of stakeholders: the data owners and the data users. Owners will customize what the users will see and the privacy-preserving rules will control and ensure potential attack scenarios are handled through appropriate checks and balances within the tool.

Visualization approaches for climate data analysis

Over the last year, as part of my postdoctoral work under Professor Claudio Silva, I have been working on the DataONE project as part of the Exploration, Visualization, and Analysis (EVA) working group. Dr. Bob Cook, a climate scientist from Oak Ridge National Lab, is leading the group.

I have closely collaborated with climate modelers, geologists and ecologists and led a team of visualization researchers, including Dr. Enrico Bertini (Assistant Professor at NYU-Poly) and Jorge Poco (PhD student at NYU-Poly). The goal of this collaboration has been to understand how visualization can best fit into the scientists' scheme of analyzing massive amounts of multifaceted, multi-scale, spatiotemporal data related to comparison of climate models.

As part of my research efforts, I have led two parallel projects in the past year. In the first project, we performed a detailed qualitative analysis of the visualizations designed by climate scientists for analyzing their shortcomings in terms of visualization design and consequently suggesting alternative solutions. Current visualization research is largely based on only assumptions and in some cases, conjectures about what people need and do, and I think this can greatly limit the impact of visualization in the future. Our study aimed to fill this gap, by using a grounded theoretical approach that led to taxonomy of design problems that we propose. This was followed by interviews for investigating the causes of matches and mismatches about design problems between climate scientists and visualization experts [11].

The outcome of the discussions was used as an incentive for the second ongoing project, the goal of which was to design an interactive visualization system for analyzing similarity of climate models [13]. The tool allows climate scientists to integrate disparate aspects of climate models in a coherent way. Our tool not only allows them to confirm existing hypotheses but also form new questions about model behavior. I consider this as a huge contribution to the state-of-the-art in the climate science community. The feedback and response that our tool has received so far has been positive and holds much promise for promise for the future. This has encouraged us to extend our current collaboration in the form of authoring efforts for procuring grants for continuing the project. We are currently exploring those avenues by jointly authoring grant proposals led by Dr. Bertini and Dr. Cook.

Visual Uncertainty Models

The information visualization literature has traditionally lacked quantitative methods for measuring the perceptual implications of what we show to the users and adapting the visual representations for maximizing user benefit. This is critical in application scenarios where the data cardinality and dimensionality far exceed the screen-space real estate.

In order to address this question, as part of my dissertation thesis, I had proposed the visual uncertainty paradigm [3, 4,] for quantifying screen-space information in visualization. The paradigm is based on a simple mantra: measure, optimize

and adapt. What we measure is the uncertainty stemming from the visualization process due to the various data transformation and mapping stages, which is different from data uncertainty.

I have demonstrated the practical utility of these metrics by applying them in the context of high-dimensional data analysis [2, 7]. One of the related projects was done in collaboration with Dr. Luke Gosink and a group of biologists from Pacific Northwest National Laboratory. As part of that project I developed a tool for analyzing high-dimensional, time-varying data by using the metrics for quantifying interesting features in the screen-space [12].

My initial results for metric-based optimization are promising and there is a lot of scope for future research. I am currently building on top of my thesis results by extending the visual uncertainty paradigm and coming up with a taxonomy that can be applied for building a comprehensive theoretical model for bridging user perception and cognition with the core visualization pipeline. Currently, I am collaborating with Dr. Anushka Anand at Tableau Software for devising metrics for finding the optimal visual mappings in the context of high-dimensional data analysis.

I believe the direction of visual uncertainty will also let us systematically evaluate visualization designs, by empirically validating the impact of common design problems in the context of scientific data analysis. In our interactions with climate scientists, we found a number of common practices, like use of rainbow color maps, or sub-optimal choice of visual variables, which are in clear conflict with visualization best practices. But we still do not have enough empirical evidence to suggest if these problems lead to serious consequences like misinterpretation or inaccuracy in estimation of the data. Teasing apart the causes and effects of these design elements (such as color maps, visual variables) in terms of visual uncertainty will eventually allow us design controlled experiments for objectively measuring the effect of these problems.

Future Research Directions

In the future, I would like to continue working with domain experts and extend my collaboration to other domains such as healthcare, finance, security, etc. More such collaborations will shed light on how visualizations are used 'in the wild' and help introspecting on the shortcomings in the current visualization literature.

A key focus area for my research will be building techniques and visualization approaches for handling heterogeneous data emerging from disparate sources such as social media, news articles, blogs, urban infrastructure etc.

I also look forward to leading research projects for testing the applicability of visual uncertainty models in visualizing complex data arising from these sources, which will help validate the benefits of the closed-loop system.
Effective application of perceptually motivated metrics is still an under-researched area in visualization. However, to meet the demands of the ever-increasing complexity of the data, we need to leverage these metrics for building visualization systems that can adapt dynamically in real-time, to users' need and interactions [18]. I believe the way forward is a bottom-up approach towards developing, applying, and validating these metrics in real-world application scenarios.

Preserving data privacy is a ubiquitous problem in the big data era and I strongly believe that novel visualization based approaches can offer alternative solutions that have not been looked into, till date.

I have so far approached visualization research with a dual focus on building upon the fundamentals of visualization, especially the perceptual and cognitive side; and on building models and techniques for real-world application scenarios. In the future I would continue to pursue both the theoretical and applied directions of visualization research, as I believe there is a symbiotic association between these two threads of research; application motivates the theoretical research, and the theory work flows back into improving the application.